Метрики моделей бинарной классификации

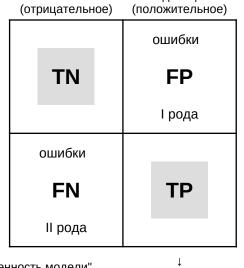
Модель говорит:

. единица

Матрица ошибок

На самом деле: ноль (отрицательное значение)

На самом деле: единица (положительное значение)



Модель говорит:

ноль

По смыслу: "ответственность модели" Доля на самом деле положительных среди предсказанных положительно, снижается с ошибками I рода (FP)

Precision

"Эпидемические" метрики: как точно модель отражает реальный мир?

Specificity (специфичность), Recall нулевого класса

Доля верно определенных отрицательных в выборке

Recall (полнота), Sensitivity (чувствительность), True Positive Rate

Доля верно определенных положительных в выборке, снижается с ошибками II рода (FN)

Будьте внимательны: порядок на осях как в sklearn, а не как в Википедии!

Подсказка к расчету метрик:

True-показатели (серый фон) — в числителе, сумма показателей на линии — в знаменателе

Дисбаланс классов. Примеры.

1. Дисбаланс классов 90/10, модель в 99% случаев предсказывает нулевой класс

	TN:	90	FP:	0	90	отрицательных в выборке
	FN:	9	TP:	1	10	положительных в выборке
		99		1	100	
	модель считает отрицательными		модель считает положительными			
Accuracy	_	TN+TP Total		0.910	0.090	Misclassification rate = (1 - Accuracy)
Recall	_	TP TP+FN		0.100	0.550	Balanced Accuracy = (Recall + Specificity) / 2
Specificity	_	TN TN+FP		1.000	0.000	False Positive Rate = (1 - Specificity)
Precision	_	TP TP+FP		1.000		
F1 Score		recision * Receision + Rec		0.182		

2. Дисбаланс классов 90/10, в каждом классе верно предсказано 80%

	TN:	72	FP:	18	90	отрицательных в выборке
	FN:	2	TP:	8	10	положительных в выборке
		74		26	100	
	модель считает отрицательными			модель считает положительными		
Accuracy	_	TN+TP Total		0.800	0.200	Misclassification rate
Recall	_	TP TP+FN		0.800	0.800	Balanced Accuracy
Specificity	_	TN TN+FP		0.800	0.200	False Positive Rate
Precision	_	TP TP+FP		0.308		
F1 Score		Precision * Recision + Rec		0.444		

Построение ROC-кривой

ROC расшифровывается (хотя, скорее зашифровывается) как Receiver Operating Characteristic.

Модели-классификаторы возвращают Score, значения в диапазоне от 0 до 1, которые удобно считать вероятностью единичного класса. Решение об отнесении измерения к единичному классу принимается в зависимости от выбранного порога threshold. Если score > threshold, то считаем, что предсказан единичный (положительный) класс.

В зависимости от порога меняются метрики модели. Для визуализации изменений используется ROC-кривая. Каждая точка кривой — это значения пары метрик модели для одного из порогов. Количество точек равно выбранному количеству порогов. Принято начинать кривую в точке (0, 0) и заканчивать в точке (1, 1).

По оси X: False Positive Rate, по оси Y: True Positive Rate

Для построения кривой нужна таблица истинных классов, отсортированная по убыванию score.

Метрика ROC_AUC вычисляется как площадь под ROC-кривой (area under curve). В идеале roc_auc = 1. Диагональная линия показывает, как в среднем работал бы классификатор на базе подбрасывания монетки.

Пример

Столбцы таблицы	Описание
Yn	Если истинный класс положительный — 0, иначе — 1
Yp	Если истинный класс положительный — 1, иначе — 0
Sc	Значения Score, полученные от классификатора
FPR	False Positive Rate, отношение текущего значения FP к максимальному в столбце
FP	Число False Positive, рассчитанное в предположении, что порог строчкой ниже
TP	Число True Positive, рассчитанное в предположении, что порог строчкой ниже
TPR	True Positive Rate, отношение текущего значения ТР к максимальному в столбце
AUC	Часть площади под кривой между предыдущей и данной точкой

Заданные значения			Рассчитанные значения						
				Ось Х			Ось Ү		Предсказанный класс
Yn	Υp	Истинный класс	Sc	FPR	FP	TP	TPR	AUC	в случае threshold = 0.5
		искусственная точка [0, 0]		0			0		
0	1	положительный	0.9	0	0	1	0.2	0	положительный
0	1	положительный	8.0	0	0	2	0.4	0	положительный
0	1	положительный	0.7	0	0	3	0.6	0	положительный
1	0	отрицательный	0.6	0.25	1	3	0.6	0.15	положительный
0	1	положительный	0.5	0.25	1	4	8.0	0	отрицательный
1	0	отрицательный	0.4	0.5	2	4	8.0	0.2	отрицательный
0	1	положительный	0.3	0.5	2	5	1	0	отрицательный
1	0	отрицательный	0.2	0.75	3	5	1	0.25	отрицательный
1	0	отрицательный	0.1	1	4	5	1	0.25	отрицательный



